Cross-language Learning with Adversarial Neural Networks: Application to Community Question Answering

Shafiq Joty, Preslav Nakov, Lluís Márquez and Israa Jaradat
Qatar Computing Research Institute, HBKU

Cross-Language Adversarial Neural Network (CLANN)

Adversarial Training

We put a language discriminator, another neural network that takes the internal representation of the network \(f\) as input, and tries to discriminate between English and Arabic \(q\).

- The discriminator is defined by another sigmoid: \(L_0 = p(l) = l(f, \theta) = \text{sigmoid}(w^T h)\)
- \(h = g(U, f, \theta, \omega)\) defines the hidden layer of the discriminator
- Discrimination loss: \(L_d = -\log L_0 - (1 - l) \log (1 - L_0)\)

Overall training objective of the composite model:

\[
L(\theta, \omega) = \frac{1}{n} \sum_{i=1}^{n} L_0(\theta) - \frac{1}{n} \sum_{i=1}^{n} L_0(\omega) + \frac{1}{n} \sum_{i=1}^{n} L_q(\omega) = \left(1 - \epsilon \right) \log (1 - \epsilon_{\omega})
\]

where \(\theta = \{U, V, w\}, \omega = \{U, V, w, U_t, w_t\}\), and \(\lambda\) controls the relative strength of the two networks.

In training, we look for parameters that satisfy a min-max optimization criterion:

\[
\theta^* = \arg \min_{\theta} \max_{\omega} \mathbb{E}_{(U, V, w)} L(\theta, \omega)
\]

The updates of the shared parameters \((U, V, w)\) for the two classifiers is done in an adversarial way.

Features

- Cross-language embeddings trained with bivec (Luong et al. 2015) to map \(q\) and \(q'\) to fixed-length vectors \(x_q\) and \(x_{q'}\)
- yields better initialization
- crucial when there is no enough labeled data to learn the input representations with end-to-end training
- Model interactions between \(x_q\) and \(x_{q'}\):
- \(\phi = g(U, f, \theta, \omega)\)
- Use pairwise features \(\phi(q, q')\) to encode similarity directly:
  - \(f = g(U, V, w, \phi(q, q'))\)
  - \(\phi(q, q')\) encode different similarity measures and task-specific features
- A non-linear transformation allows us to learn high-level abstract features based on these pairwise features.

- The classification layer is defined by a sigmoid:
  - \(\epsilon = p(c = 1|f, \phi(q, q')) = \text{sigmoid}(w^T f(\phi(q, q'))\)
- We optimize the log probability:
  - \(\mathcal{L}(\phi) = -\log \epsilon + (1 - \epsilon) \log (1 - \epsilon)\)
- This network learns features that are discriminative for the classification task, i.e., similar vs. non-similar. However, our goal is also to make these features invariant across languages.

Dataset

Based on the SemEval-2016 Task 3 dataset
- 387 original questions (276.50, and 70 for training, development and test)
- For each original question 10 related questions to be ranked
- We translated the 387 original questions manually to Arabic.
- We further collected 221 original and 1,863 related questions (English; unlabeled). We manually translated the 221 questions to Arabic.

Unsupervised Adaptation

<table>
<thead>
<tr>
<th>System</th>
<th>Discrim.</th>
<th>MAP</th>
<th>MRR</th>
<th>AvgRec</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNN</td>
<td>en→ar</td>
<td>75.28</td>
<td>84.26</td>
<td>89.48</td>
</tr>
<tr>
<td>CLANN</td>
<td>en→ar</td>
<td>76.64</td>
<td>84.52</td>
<td>90.92</td>
</tr>
<tr>
<td>FNN</td>
<td>ar→en</td>
<td>75.32</td>
<td>84.17</td>
<td>89.26</td>
</tr>
<tr>
<td>CLANN</td>
<td>ar→en</td>
<td>76.70</td>
<td>84.52</td>
<td>90.61</td>
</tr>
</tbody>
</table>

Semi-supervised Adaptation

<table>
<thead>
<tr>
<th>System</th>
<th>Discrim.</th>
<th>MAP</th>
<th>MRR</th>
<th>AvgRec</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNN</td>
<td>en→ar</td>
<td>74.69</td>
<td>81.79</td>
<td>88.15</td>
</tr>
<tr>
<td>CLANN</td>
<td>en→ar</td>
<td>75.93</td>
<td>84.15</td>
<td>89.63</td>
</tr>
<tr>
<td>CLANN</td>
<td>en+ar*→ar</td>
<td>76.65</td>
<td>84.52</td>
<td>90.84</td>
</tr>
</tbody>
</table>

Unsupervised Adaptation

Visualizing the Representation Layer

Arabic=blue, English=red. Class labels \{0,1\}. L: ar→en, R: en→ar

Conclusion

We have studied cross-language adaptation for question-question similarity in community question answering, in order to port a system trained on one input language to another input language. This is novel in a cross-language setting.

Future work

- Fine-tune the word embeddings for the cross-language task
- Try LSTM and CNN
- Experiment with more than two languages at a time
- Apply to other tasks